Early Detection and Forecasting of Research Trends

Angelo Antonio Salatino

Knowledge Media Institute, The Open University, United Kingdom
angelo.salatino@open.ac.uk

Abstract. Identifying and forecasting research trends is of critical importance for a variety of stakeholders, including researchers, academic publishers, institutional funding bodies, companies operating in the innovation space and others. Currently, this task is performed either by domain experts, with the assistance of tools for exploring research data, or by automatic approaches. The constant increase of research data makes the second solution more appropriate, however automatic methods suffer from a number of limitations. For instance, they are unable to detect emerging but yet unlabelled research areas (e.g., Semantic Web before 2000). Furthermore, they usually quantify the popularity of a topic simply in terms of the number of related publications or authors for each year; hence they can provide good forecasts only on trends which have existed for at least 3-4 years. This doctoral work aims at solving these limitations by providing a novel approach for the early detection and forecasting of research trends that will take advantage of the rich variety of semantic relationships between research entities (e.g., authors, workshops, communities) and of social media data (e.g., tweets, blogs).

Keywords: Scholarly Data, Research Trends, Trend Detection, Trend Forecasting, Semantic Web Technologies.

1 Problem Statement

The research environment evolves rapidly: new potentially interesting research areas emerge regularly while others fade out, making it difficult to keep up with such dynamics. The ability to recognise important new trends in research and forecasting their future impact is however critical not just for obvious stakeholders, such as researchers, institutional funding bodies, academic publishers, and companies operating in the innovation space, but also for any organization whose survival and prosperity depends on its ability to remain at the forefront of innovation.

Currently, the task of understanding what the main emergent research areas are and estimating their potential is usually accomplished by experts with the help of a number of systems for making sense of research data. Systems such as Google Scholar, FacetedDBLP [1] and CiteSeerX [2] provide good interfaces which allow users to find scientific papers, but they do not directly support identification of research trends. Other tools such as Microsoft Academic Search, Rexplore [3], Arnetminer [4], and Saffron [5] provide a variety of visualizations that can be used for trend analysis, such as publication trends and co-authorship paths among researchers. However, the manual detection of research trends is an intensive and time-consuming task. Moreover, the constant increase in the number of research data published every year makes the approach based on human experts less and less feasible. It is thus important to develop automatic and scalable methods to detect emerging research trends and estimate their future impact.

Currently, there are a number of approaches for detecting topic trends in a fully automatic way [6,7]. These are usually based on the statistical analysis of the impact of certain labels associated with a topic. However, these tools are unable to take full advantage of the variety of research data existing today and need to examine a significant number of years (e.g., 3-4) before they are able to identify and forecast topic trends [8,9]. In addition, they are only able to identify topics that have been explicitly labelled and recognized by researchers [10]. However, it can be argued that a number of topics start to exist in an embryonic way, often as a combination of other topics, before being officially named by researchers. For example, the Semantic Web emerged as a common area for researchers working on Artificial Intelligence, WWW and Knowledge-Based Systems, before being recognized and labelled in the 2001 paper by Tim Berners-Lee et al. [11].

The doctoral work presented here aims to solve the aforementioned limitations and produce a novel approach to detect and forecast research topics. This approach will be based on two main intuitions. First, I believe that by analysing the various dynamics of research it should be possible to detect a number of patterns that are correlated with the creation of new embryonic topics, not yet labelled. For example, the fact that a number of authors from previously unrelated research communities or topics are starting to collaborate together may suggest the emergence of a new interdisciplinary research area. Secondly, I theorize that taking into account the rich variety of semantic relationships between research entities (e.g., authors, workshops and communities) and analysing their diachronic evolution, it should become possible to forecast a topic impact in a much shorter timescale, e.g., 6-18 months. This holistic and semantic-based analysis of the research environment is today made possible by the abundance of both scholarly data and other sources of evidence about research, including social networks, blogs, and so on.

2 Relevancy

In many real-world contexts, being aware of research dynamics can bring significant benefits. **Researchers** need to be updated regularly on the evolution of research environments because they are interested in new trends related to their topics and potentially interesting new research areas. For **academic publishers** or **editors** knowing in advance new emerging topics is crucial for offering the most up to date and interesting contents. For example, an editor can gain a competitive advantage by being the first one to recognize the importance of a new trend and publish a special issue or a journal about it. **Institutional funding bodies** and **companies** need also to be aware of research developments and promising research trends. Thus, an automatic approach to detect novel topics and estimate their potential will bring significant advantages to a variety of stakeholders. Indeed support for this PhD project comes from Springer-Verlag, which is a global publishing company.

3 Related work

Several tools and approaches for the exploration of scholarly data already exist. From the perspective of topic trend detection, we can classify these systems as either semi-automatic or fully automatic. In particular, some systems for exploring the publication space provide implicit support for semi-automatic trend detection, such as Google Scholar, FacetedDBLP [1] and CiteSeerX [2]. Other systems offer instead an explicit support for semi-automatic trend detection, like Arnetminer [4], Microsoft Academic Search (MAS), Saffron [5] and Rexplore [3]. However, while all these systems are able to identify and visualize historical research trends, they do not provide any support for the detection of future ones.

In the context of providing a fully-automatic way to detecting topic trends, many approaches assess the impact of a topic by simply using the number of publications or patents directly associated with it. For example, Wu et al [8] integrate bibliometric analysis, patent analysis and text-mining analysis in order to detect research trends. Some models also take in consideration the citation graph. For example, Bolelli et al. [6] propose an author-topic model to identify topic evolution and then they use citations to evaluate the weight for the main terms in documents. He et al. [7] combine Latent Dirichlet Allocation and citation networks for detecting topics and understand their evolution. However, these approaches are able to detect trends only after the associated research areas are already established and they do not provide any support to the early detection of research trends.

State of the art methods for forecasting trends in research take usually into consideration the number of publications and authors associated with a topic [12], or the probability distribution of a topic over time [13]. They then analyse these time series either by means of statistical techniques [10] or machine learning methods [14], yielding a prediction for the following years. However, these methods do not take advantage of the knowledge that can be extracted by analysing the dynamics of multiple research entities (e.g., communities, venues), and they ignore the growing mass of research data that today can be acquired from social networks.

Another important aspect that needs to be taken into account is how to represent a topic. In literature, several ways to define a topic model can be found. The first is characterised by the use of keywords as proxies for research topics. Systems like MAS and Saffron [5] use this kind of model. This approach has several drawbacks because it does not take in consideration the relationships among research topics [15] and keywords tend to be noisy. The second kind of approach is the probabilistic topic model. Latent Dirichlet Allocation [16], which treats a document as a mixture of topics and a topic as a distribution over words, is the most popular of these methods. However, this model assumes that the topics used to generate a document are uncorrelated, which may be a risky assumption for research topics [17]. Other approaches for probabilistic topic model try to deal with this problem introducing a separability condition [18]. A third solution is using an explicit semantic topic model [9,17,3], which exploits a semantic network of research areas linked by semantic relations. The advantage of this solution is that it goes beyond the use of noisy, uncorrelated keywords and exploits instead an ontology of research areas.

4 Research Questions

Considering the gaps identified in the previous section, the main research question of the PhD will be: *"How is it possible to detect the early emergence of new research topics and forecast their future impact?"*.

This question entails two different challenges. The first one is how to detect very early research topics that may not even be labelled. The second one is how to forecast their impact with good accuracy. A specific set of sub-questions has been articulated in order to describe the process through which the doctoral work plans to answer the questions above.

Q1 – Finding the data. Understanding which data to integrate and exploit for the process is the first step. In particular, it is important to investigate the value of non-scholarly data (e.g., tweets, blogs, micro-posts, slides) in supporting trend detection and forecasting. As far as semantic technologies are concerned: how can research elements be gathered and connected by means of semantic relations?

Q2 – Detection of new emerging research topic. How is it possible to extract patterns in the evolution of research areas in order to predict the emergences of new ones? How can historical patterns be used to support the detection of future trends? Is it possible to develop a general approach able to consider the peculiarities of different fields (e.g., Computer Science, Business, Medicine and so on)? How emerging and unnamed research areas can be labelled? How social media can contribute in the detection of research trends?

Q3 – Forecasting of research trends. Can the impact of a research topic be measured just in terms of number of citations and publications? As soon as it has been defined, how the impact of research areas can be forecasted? What kind of forecasting approach should be adopted for research areas that do not yet exist? Which contribution can be given from the social media?

5 Hypotheses

From a philosophical point of view, academic disciplines are specific branches of knowledge which together form the unity of knowledge that has been produced by the scientific endeavour. When two or more disciplines start to cooperate they share their theories, concepts, methods and tools. The results of this cooperation may lead either to the creation of a new interdisciplinary research area or simply to a contribution in knowledge from one area to another. The basic hypothesis is that the creation of a topic is thus anticipated by a number of dynamics involving a variety of research entities, such as other topics, research communities, authors, venues and so on. Therefore, recognizing these dynamics might enable a very early detection of emerging topics.

Scholarly data can be used to analyse a huge amount of research elements such as papers, authors, affiliations, venues, topic and communities [19]. All these research elements are inherently interconnected by relations that can be defined as either explicit or implicit. Figure 1 shows, as an example, the six basic explicit connections between the research elements according to our model.

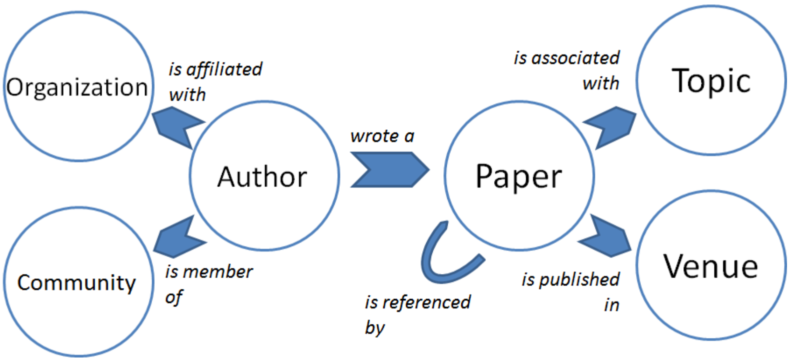

Fig. 1. Model representing the scholarly meta-data and their relationships

These explicit connections can be used to derive a number of second order connections, e.g. a topic is also associated with publication venues through relevant papers published in venues. These relationships can be analysed diachronically to derive the dynamics that led to the emergence of a topic and to estimate how they affect its future impact. For example, if two communities start to share research interests or authors, this may lead to the fact that a common new topic is developing. In a nutshell, the fundamental hypothesis at the basis of this PhD is that by exploiting the large variety of scholarly data which are now available, as well as modelling their semantic relationships, it will be possible to perform detection and forecasting of research trends even in a relative small interval of time. In addition, since many researchers are actively involved on social networks, I believe that analysing data from social media can also provide an effective support for the detection of research dynamics.

6 Approach

The approach is structured according to the proposed research questions. Basically, it is organised in four main steps.

Data integration. In this first phase I plan to integrate a variety of heterogeneous data sources, including both scholarly metadata and less traditional sources of knowledge, such as tweets, blogs post, slides and so on. The output will be a comprehensive knowledge base containing both the research entities from Figure 1 and entities from social media (authors' profiles, number of followers, analytics, etc.). I will identify topics and communities by extending state of the art techniques. In particular, I plan to treat topics semantically, by describing their relationships using the topic networks produced by the Klink algorithm [17]. I am also planning to use the approach for detecting topic-based research communities described in [19], since it explicitly links communities and topics.

The rich network of semantic relationship between the research elements will be described by an ontology and it will be populated by semi-automatic statistical methods. To build it, I plan to extend the topic network created by Klink with the research entities discussed in section 5 and their relationships. The analysis of these relationships and how they change in time will support the next steps of the approach.

Exploration of the Research Dynamics. In this step, the dynamics involving research elements correlated with the emergence of new topics will be investigated. To do so, I plan to verify empirically a number of hypotheses about these dynamics. In particular, I will analyse a number of topics which appear in the 2000-2010 interval and verify if their emergence is correlated with a number of dynamics, such as the raise of co-publications of related research areas, the increase of collaborations between authors of related areas, shifts of interests or migration phenomena in related communities, transfer of topics between related venues, and so on.

The output of this analysis will be a collection of patterns of knowledge flows associated with the creation of a new research area.

Early topic detection. This step aims to exploit the previously defined patterns for early research trend detection. To this end, I will build a number of distinct graphs, in which nodes represent a kind of research entity (e.g., topics) and the links are one of the elements of the dynamics, which were found in the previous phase – e.g., the increase in the number of collaborations between authors from two distinct topics. Highly connected sub-graphs, representing the area in which multiple entities exhibit the identified dynamics could thus suggest that a new discipline is emerging. In order to produce more robust evidence, I will use the semantic network of research entities to confirm that the emergence of a new topic is supported by a number of different 'traces' and research entities. For example, if a set of topics suggests that a correlated research area is emerging, the dynamics of the set of communities and venues related to these topics will also be checked. The intuition is that, while the evidence coming from a single dynamics or a single kind of entity could be biased or noisy, their combination should yield a more accurate result. The result will be a number of sets of linked entities, each one anticipating the emergence of a new topic. Different kinds of combination of entities and metrics will be tested, aiming to find the best approach to derive sets that are strongly correlated with the creation of new topics. At this stage, another challenge will be the definition of a method for labelling future research topics.

Trend forecasting. Initially, I will investigate different techniques to estimate the impact of a topic, taking in consideration both basic metrics, such as the number of publications and citations, and more complex indexes. As mentioned before, in contrast with current approaches, [8,9], I aim to develop a method which will be able to work also on relatively short time series (6-18 months). In order to do so, I will take advantage of a wide variety of features associated with a topic, representing both the performances of related entities (e.g., the track record of significant authors) and the previously discussed dynamics. Hence, I will conduct a comprehensive analysis of the correlations between these features and the topic impact in the following years. For example, I will analyse how the performance of related authors, communities, workshops, hashtags, scientific opinion leaders, and so on, influence on the previously defined impact metrics. It is hypothesised that such abundance and diversity of the features will compensate for the small interval of time in which early topics will be analysed. Moreover data from the social web and other real-time information, such as the number of views and downloads on the publisher sites and open access repositories, will offer a more granular timeline for the analysis of the topics, measured in weeks, rather than in years.

A set of different machine learning methods, such as Artificial Neural Networks, Support Vector Machines and Deep Belief Networks, will exploit the extracted features in order to forecast the performance of a topic.

7 Evaluation plan

I plan to conduct an iterative evaluation during the different phases of my work using both quantitative and qualitative approaches.

From a quantitative point of view, I will evaluate both the ability of the system to identify novel topics and its accuracy to assess their impact in the following years. The discussed approaches will be compared with current methods and the difference between their performances will be measured via statistical tests. I will evaluate the detection of emerging trends in terms of recall, precision and F-measure using cross-validation on historical data. Similarly, I will assess the agreement between the estimated and the real impact of a research area.

In the qualitative evaluation, the achieved results will be compared with experts' opinions in order to measure its reliability. I will prepare a number of surveys for domain experts with questions both about the past - such as the main topics recently emerged in their area of expertise - and about the future - such as the research areas which seem on the verge of being created and an estimation of their likely impact.

8 Conclusions

This paper presents the goal of my doctoral work, which is currently at an early stage (month 6). As discussed, I intend to produce a new approach for detecting and forecasting research trends, which is based on a semantic characterization of research entities, on the statistical analysis of research dynamics and on the integration of scholarly and social media data.

Currently I am investigating a number of knowledge sources for selecting the ones more apt to support my approach. At the same time I am using an initial dataset to test the hypotheses about research dynamics discussed in section 6. The next step will be the creation of an approach for extracting highly connected sub-graphs of entities exhibiting dynamics associated with the emergence of new topics.

Acknowledgements

I would like to acknowledge my advisors, Enrico Motta and Francesco Osborne, as well as my sponsors, Springer-Verlag GMBH for supporting my research.

References

1. Diederich J, Balke W-T, Thaden U Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 2007. ACM, pp 505-505

2. Li H, Councill I, Lee W-C, Giles CL CiteSeerx: an architecture and web service design for an academic document search engine. In: Proceedings of the 15th international conference on World Wide Web, 2006. ACM, pp 883-884

3. Osborne F, Motta E Rexplore: Unveiling the dynamics of scholarly data. In: Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on, 2014. IEEE, pp 415-416

4. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD, 2008. ACM, pp 990-998

5. Monaghan F, Bordea G, Samp K, Buitelaar P Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food. In: SW Challenge - ISWC, 2010. Citeseer, pp 420-435

6. Bolelli L, Ertekin Ş, Giles CL (2009) Topic and trend detection in text collections using latent dirichlet allocation. In: Advances in Information Retrieval. Springer, pp 776-780

7. He Q, Chen B, Pei J, Qiu B, Mitra P, Giles L Detecting topic evolution in scientific literature: how can citations help? In: Proceedings of the 18th CIKM, 2009. ACM, pp 957-966

8. Wu F-S, Hsu C-C, Lee P-C, Su H-N (2011) A systematic approach for integrated trend analysis—The case of etching. Technological Forecasting and Social Change 78 (3):386-407

9. Decker SL, Aleman-Meza B, Cameron D, Arpinar IB (2007) Detection of bursty and emerging trends towards identification of researchers at the early stage of trends. (Doctoral dissertation, University of Georgia).

10. Tseng Y-H, Lin Y-I, Lee Y-Y, Hung W-C, Lee C-H (2009) A comparison of methods for detecting hot topics. Scientometrics 81 (1):73-90

11. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Scientific american 284 (5):28-37

12. Budi I, Aji RF, Widodo A (2013) Prediction of Research Topics on Science & Technology (S&T) using Ensemble Forecasting. International Journal of Software Engineering and Its Applications 7 (5):253-268

13. Zhou D, Ji X, Zha H, Giles CL Topic evolution and social interactions: how authors effect research. In: Proceedings of the 15th CIKM '06, 2006. ACM, pp 248-257

14. Jun S, Uhm D (2010) Technology forecasting using frequency time series model: Bio-technology patent analysis. Journal of Modern Mathematics and Statistics 4 (3):101-104

15. Osborne F, Motta E, Mulholland P (2013) Exploring scholarly data with rexplore. In: The Semantic Web–ISWC 2013. Springer, pp 460-477

16. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. the Journal of machine Learning research 3:993-1022

17. Osborne F, Motta E (2012) Mining semantic relations between research areas. The Semantic Web-ISWC 2012, pp 410-426

18. Arora S, Ge R, Halpern Y, Mimno D, Moitra A, Sontag D, Wu Y, Zhu M (2012) A practical algorithm for topic modeling with provable guarantees.

19. Osborne F, Scavo G, Motta E (2014) Identifying diachronic topic-based research communities by clustering shared research trajectories. In: The Semantic Web: Trends and Challenges. Springer, pp 114-129